DOCUMENT RESUME

ED 257 873 TM 850 340

AUTHOR Echternacht, Gary
TITLE Analyzing and Reporting School District Standardized Test Results.
PUB DATE Apr 85
NOTE 26p.; Paper presented at the National Council on Measurement in Education (Chicago, IL, April 1-4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Data Analysis; Effect Size; Elementary Secondary Education; Mathematical Models; School Districts; *Standardized Tests; Testing Programs; *Test Interpretation; *Test Results
IDENTIFIERS Tukey (John)

ABSTRACT

This paper outlines a methodology that school districts can use to enhance the presentation of district-wide testing program results to administrators, school boards, teachers, and the public. Based on John Tukey's two way analysis methodology, it involves fitting this model: test score equals overall plus year plus grade plus cohort plus residual. The method is illustrated by working through the calculations for a small district with six years of test score data in math (mean standard scores for grades 1 through 6 on the Comprehensive Test of Basic Skills). Both graphs and tables are used to summarize effect sizes. The procedure can be used in three- and four-way analyses and for percentile as well as scaled scores. The methodology is also applied to data from a large California school district to illustrate its advantages for presenting clear and meaningful summary results. Although computer technology is indispensable for many aspects of data analysis, these paper and pencil calculations can be done quickly in consulting situations and provide a familiarity with the data which can enable investigative insights. (BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED257873

ANALYZING AND REPORTING SCHOOL DISTRICT STANDARDIZED TEST RESULTS

Gary Echternacht
Educational Testing Service

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. Echternacht

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 850 340

Paper presented at the National Council for Measurement in Education's Annual Meeting, Chicago, 1985.

Analyzing and Reporting School District Standardized Test Results

Gary Echternacht
Educational Testing Service

Background

In the northeast most school districts have district-wide testing programs where they test all students in reading and math at each of the elementary school grades and sometimes through the high school grades. This testing program must serve many masters. It acts as an accountability indicator for the schools and the school district. It is used to evaluate school programs and to identify students who need special help.

School district testing coordinators routinely present the results of the district-wide testing to administrators, the school board, teachers and the public. In large school systems, lengthy reports are prepared giving the results for the district and each school. In smaller districts, shorter reports or oral presentations are made to the board, administrators and the public alike.

The purpose of this paper is to outline a methodology that school districts can use to enhance presentation of the results of the testing. The methodology has been found useful by a number of small to medium sized school districts in

New York and New Jersey. I would like to say that these districts routinely include the analyses in their reports, but that has yet to happen. School people are a conservative lot and their reports look much the same from year to year. Nevertheless, a number of school district test coordinators perform the analyses on their own, to better understand what information is present in the data and what the data suggest is occurring with regard to achievement.

The Method

Generally, a report on a district-wide testing program provides statistics that are useful in answering three questions:

- Is achievement in the district improving?
- How do different schools in the district compare with regard to achievement?
- How are students progressing as they move through the school system?

Data relating to these questions may take the form illustrated below.

Mean Standard Scores in Reading (CTBS)

Grade	1982	1981	1980	1979	1978	1977
1	352	356	345	362	345	314
2	415	406	424	421	407	381
3	456	455	451	461	453	453
4	513	508	478	486	485	483
5	546	532	535	533	530	524
6	562	563	571	569	560	539

The above is a two-way table. It shows a trend in scores from 1977 to 1982. Similarly, it shows trends from grade one to six. Unfortunately, it shows these trends with 36 numbers, which makes it difficult to draw any conclusions. What we need is a way to simplify such tables—i.e., a way to break them down into component parts.

John Tukey* has developed a methodology that enables one to perform two-way analyses in a smiple and straightforward way. It involves fitting the model:

$$\text{Data} = \text{overall} + \text{row} + \text{column} + \text{residual}$$

In our example above, a row effect would be equivalent to a grade level effect, while a column effect would be equivalent to a year effect. In analyzing data of the sort above, I have found it useful to add a third effect, making the model:

$$\begin{matrix}\text{test}\\ \text{score}\end{matrix} = \text{overall} + \text{year} + \text{grade} + \text{cohort} + \text{residual}$$

Cohort is a term that accounts for differences in achievement between groups as they pass through the school system. It is the extent, for example, that last year's sixth graders were better achieving than this year's sixth graders. Although the

---

*Tukey, J. Exploratory Data Analysis. Reading, Ma. Addison-Wessley. 1977. pp. 331-600.

composition of cohorts change from year to year in many districts the change is so slight that it can be ignored. That we have found convenient to do.

Two-way Analyses

We will illustrate the method by working through the calculations for a district with six years of test score data in math. Mean standard scores in total math are presented by grade level.

Mean Standard Scores in Math (CTBS)

Grade	1982	1981	1980	1979	1978	1977
1	320	314	308	319	313	304
2	373	361	377	374	361	346
3	413	429	435	422	412	412
4	459	457	459	464	451	443
5	524	515	522	508	512	507
6	526	530	536	540	537	515

Details of the calculations are given on the next page. In the upper left corner of the page the above data are recorded. The single vertical line separates the row medians from the data. We then subtract the row medians from each row, placing our answer to the right of the double vertical lines. Throughout the calculation, we use Tukey's convention of using single lines to separate values from some kind of summary and double lines between things that add up. Thus, the first row values and their median is:

320 314 308 319 313 304 | 314

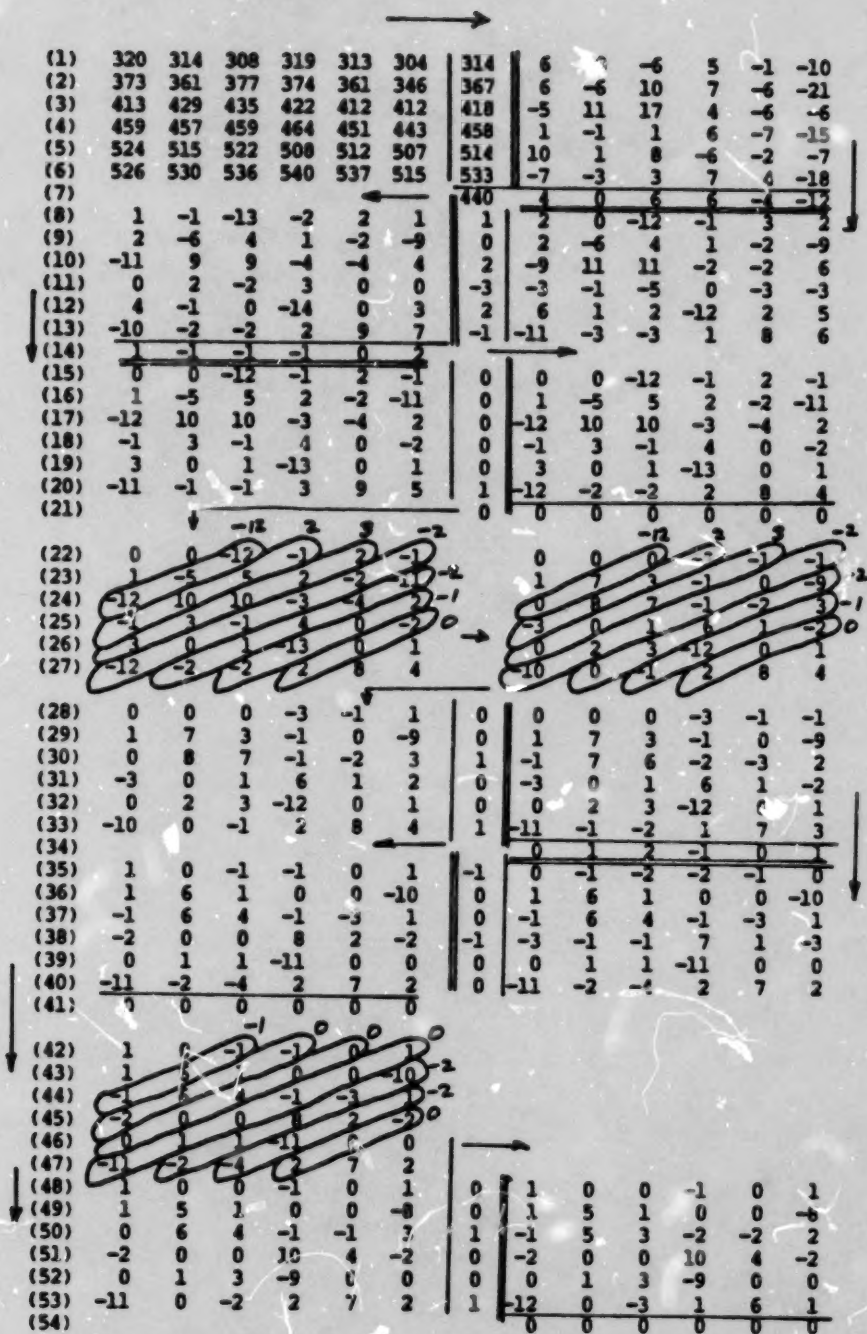
with a single line, while after subtracting, the median is:

320 314 308 319 313 304 | 314 ‖ 6 0 -6 5 -1 -10

with a double line between 314 and the differences. The first six rows continue this type calculation. Numbers on the seventh row indicate the medians of the columns. For the residuals on the right-hand side of the first six rows, the median is subtracted and placed in rows 8-13 under the residuals. The medians for the rows are placed under the first set of row medians and, in turn, are subtracted with the residuals placed in the left part of rows 8-13. The residual column medians appear on row 14. The left side of rows 15-20 show the residuals from subtracting the row 14 medians. The row medians are recorded and subtracted, giving the residuals on the right-hand side of rows 15-20. The column medians at this point gives all zeros.

In doing our calculations we have made things easy for us. We have used only whole numbers and have chosen zero for the median when choosing points midway between, say, -3 and 2 or 0 and 1.

Thus far the methodology is exactly that of Tukey's two-way analysis: row PLUS column analysis. If we look at the residuals, however, there seems to be more information. For example, five of the six main diagonal residuals are negative. The main



Column Estimates

4	0	6	6	-4	-12
1	-1	-1	-1	0	2
0	1	2	-1	0	1
5	0	7	4	-4	-9

Cohort Estimates

-12	2	3	-2	-2	-1	0
-1	0	0	0	-2	-2	0
-13	2	3	-2	-4	-3	0

Row Estimates

314 - 440 + 1 + 0 - 1 + 0 = -126
367 - 440 + 0 + 0 + 0 + 0 = -73
418 - 440 + 2 + 0 + 1 + 1 = -18
458 - 440 - 3 + 0 - 1 + 0 = 14
514 - 440 + 2 + 0 + 0 + 0 = 76
533 - 440 - 1 + 1 + 1 + 1 = 95

diagonal consists of scores earned predominantly by students who graduated from the sixth grade in 1982. On the left half of lines 22-27 we indicate cohorts where there are at least three score averages by circling the residuals from lines 15-20 and writing the median residual above or to the right of the circled data. On the right half of lines 22-27, we subtract these cohort medians.

We now go back to the residuals and adjust the row and column medians by subtraction exactly as we did before, continuing the process until zeros appear. On lines 42-47 we adjust the cohort estimates by again finding the cohort medians, subtracting them from the residuals and, this time, placing them directly on the left side of rows 48-54. Repeating the rows and column subtraction brings us to the end of the calculation.

Overall estimates are obtained by adding all the row, column, and cohort medians. The initial median of the rows must be subtracted as indicated on the page following the calculation. The final analysis can be displayed as:

Grade		1982	1981	1980	1979	1978	1977	
				-13	2	3	-2	
1	-126	1	0	0	-1	0	1	
2	-73	1	5	1	0	0	-8	-2
3	-18	-1	5	3	-2	-2	2	-4
4	14	-2	0	0	10	4	-2	0
5	76	0	1	3	-9	0	0	
6	95	-12	0	-3	1	6	1	
	440	5	0	7	4	-4	-9	

The various effects fit our model as we can verify for the 1982 third graders.

Mean = overall + grade + year + cohort + residual

413 = 440 - 18 + 5 - 13 - 1

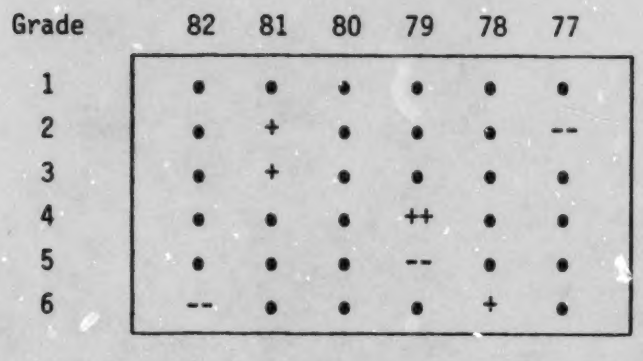
The model seems to predict score averages fairly well. It underpredicts (indicated by a large positive residual grade 2 (81), grade 3 (81), grade 6 (78), and especially grade 4 (79). It overpredicts grade 6 (82), grade 5 (79), and grade 2 (77). We depict the residuals using the following scheme:

++ large positive

\+ positive

• small positive or negative

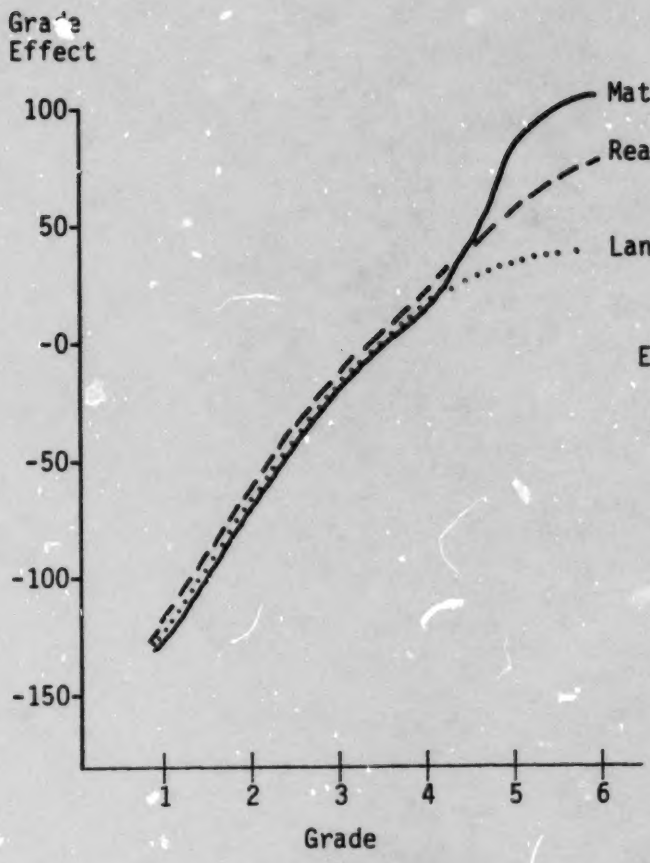
-- large negative

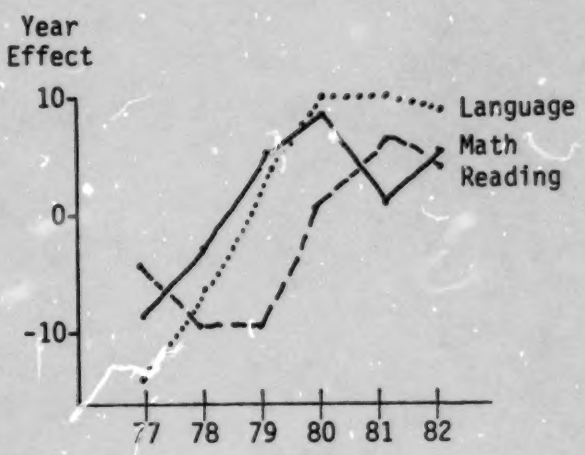


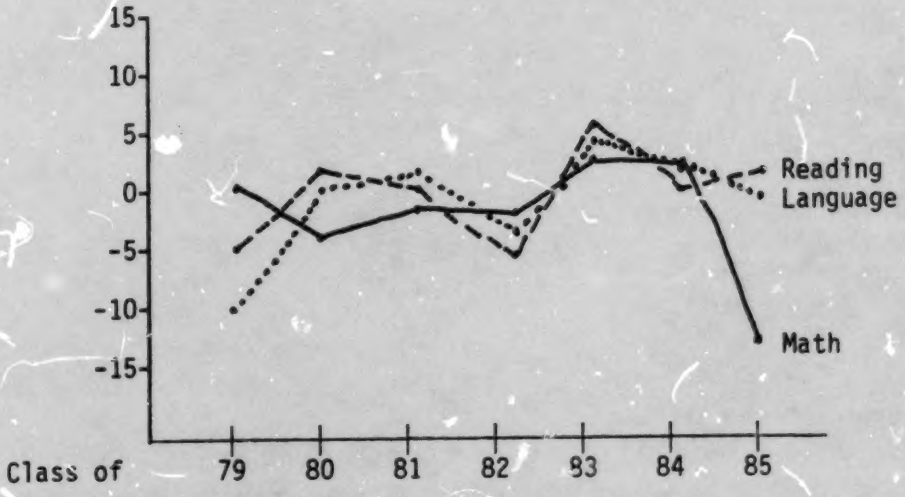
Although analysts are interested in the residuals, most people are interested in the main effects. The effects are graphed on the following page along with those from the analysis of the district's reading and language test scores.

As indicated by the size of the effects, most of the variation is accounted for by the grade level effect. Growth in reading, language, and math runs parallel up through the fourth grade. In fifth grade the effects differ. Math scores spurt higher then level off. Reading and language show more regular growth, with language tailing off quicker than reading.

Examination of the year effects show a similar pattern over each subject. Generally, scores increased between 1977 and 1980. Since 1980 they have remained high although some variations are noted, particularly the drop in 1981 for math.


Grade Effect
100
50
-0
-50
-100
-150
Math
Reading
Language
1
2
3
4
5
6
Grade


Year Effect
10
0
-10
Language
Math
Reading
77
78
79
80
81
82


15
10
5
0
-5
-10
-15
Reading
Language
Math
Class of
79
80
81
82
83
84
85

Cohort effects also appear similar. There appears evidence indicating that the class of 1982 was lower in achievement than the other classes. Similarly, the classes of 1983 and 1984 appear better in achievement than the others. Mixed evidence is found for the other classes. Both the results for the classes of 1979 and 1985 appear more variable, probably due to the small number of data elements included in the cohort effect estimates.

The same things we have done for two-way tables can be done for three- or four-way tables. In the case of the district whose data we have analyzed, they were extremely interested in comparing the overall achievement of their five elementary schools. The topic was hot, as the building principals regarded such comparisons as indicators of their effectiveness. Tukey provides a methodology for obtaining three-way analyses which is the same as that outlined above. We do not try to fit cohort estimates, however. The procedure is to construct a two-way table from the three-way table by combining two factors. Two-way analyses are performed for each level of the single factor and finally single factor effects are removed. Our purpose is to illustrate the summary rather than the methodology.

In our example, we looked at school-year combinations over grades. The overall table showing the school, year and grade

effects, and the interactions is quite large and not given. Instead, the summary table is given below.

Grade, Year and School Effects
With Second Order Interactions (Math)

Grade							
1	2	3	4	5	6	Eff	
							Year
-	-	4	-	-	-	5	82
-	-2	-	-2	2	-	1	81
3	-	-	4	-	3	6	80
-	-	-	-	-	-	4	79
-	2	-	4	-	-3	-5	78
-	-	-	-	-	-	-9	77
							School
-	-	-	-	6	-	1	Hewitt
4	-	-	-5	-	-	1	Wilson
-	3	-	-	-	-	0	Covert
-	-	-	3	-2	-	0	Riverside
-2	-	-	-	-	4	-1	Watson
-124	-73	-16	14	72	91	438	

The grade and year effects are similar to those of the previous example. The school effects are practically nonexistent. The entries in the grade-year section of the table show the grade x year interactions. Similarly, the entries in the grade-school section of the table show the grade x school interactions.

The analysis is most useful in interpreting the relative size of the effects. For example, it is clear that the largest effect is for grade level. Achievement increases largely as students mature and progress through the school system. Social and institutional forces show themselves through yearly effects. We see that achievement has increased over the six-year period with the largest increase occurring between 1978 and 1979. Recent achievement has leveled though it remains high. School effects are miniscule. For all practical purposes, they do not exist. In the vernacular, there is no reason to believe one school is any better than another.

The above analyses used scaled scores, primarily because that is the way the district reported its scores. Most districts, however, use either percentiles or grade equivalents in reporting their scores. The procedures remain the same, though the tables look somewhat different. Included below is a simple analysis of district percentile scores.

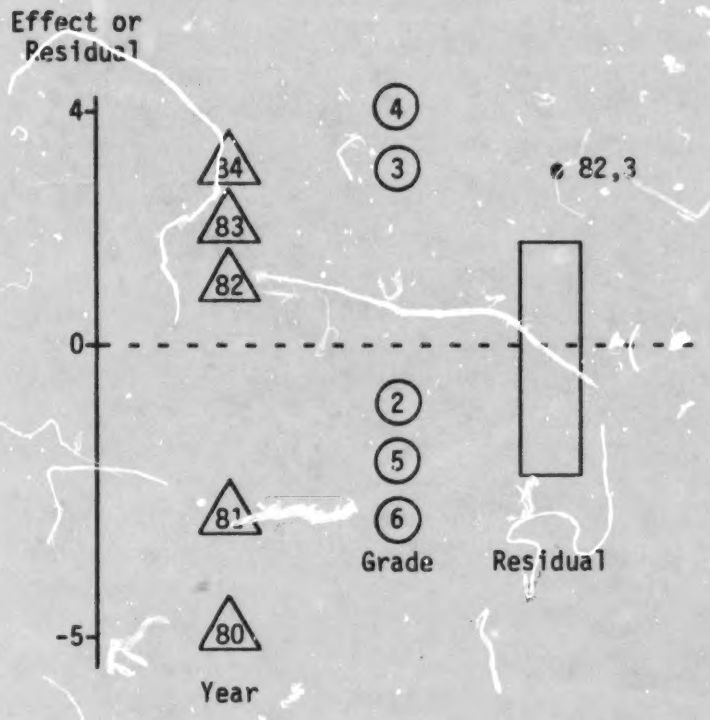
Two-way Table Breakdown of Total
Reading Percentile Rank Scores

	Year					
Grade	84	83	82	81	80	Effect
2	-2	0	0	0	2	-1
3	0	-2	3	0	0	3
4	1	1	0	-1	0	4
5	1	-1	-2	0	1	-2
6	0	1	1	-2	-2	-3
	3	2	1	-3	-5	68

In the above we see that grade effects are about the same size as year effects. Over the five years, we see that achievement has consistently improved. Achievement is about eight percentile points higher in 1984 than it was in 1980. Relative to the national norms, the district seems to do best in grades three and four in reading. The results can also be summarized by the graphic on the following page.

Although the examples presented thus far have come from consulting sessions, I thought I would include another example from a large California district, having learned that the discussant for this paper session comes from a large California

Summary of the Percentile Rank Analysis



district. The chart on the following page is presented in the district report. The report itself emphasizes district performance and ranges of performance. There is extensive data on school building performance, however.

The chart gives us little insight as it stands. There are obvious differences in achievement among the schools. Other than that, it is difficult to glean any insights.

If we use our two-way table methodology and do a little rearranging, the data suggests more. Instead of subtracting medians this time, we subtract means. With this type of data there is no cohort effect. With subtracting means, it also takes only one iteration. This is an advantage with larger data sets. We also subtract 50 from the school effects and use it as the overall statistic, primarily because it is the national average or "at grade level." The data is presented on the page following the original data.

By using 50 as the overall mean, we can represent the school effects as typical points above or below the national average. For example, without knowing the subject matter, we would say that Gompers scores about thirty-three percentile points above the national average. Lincoln, on the other hand,

DISTRICTWIDE TESTING RESULTS
COMPREHENSIVE TESTS OF BASIC SKILLS, FORM U, LEVEL K
GRADE 11 - NOVEMBER 1983
PLACENTILE RANK OF MEDIAN PUPIL BASED ON NATIONAL NORMS

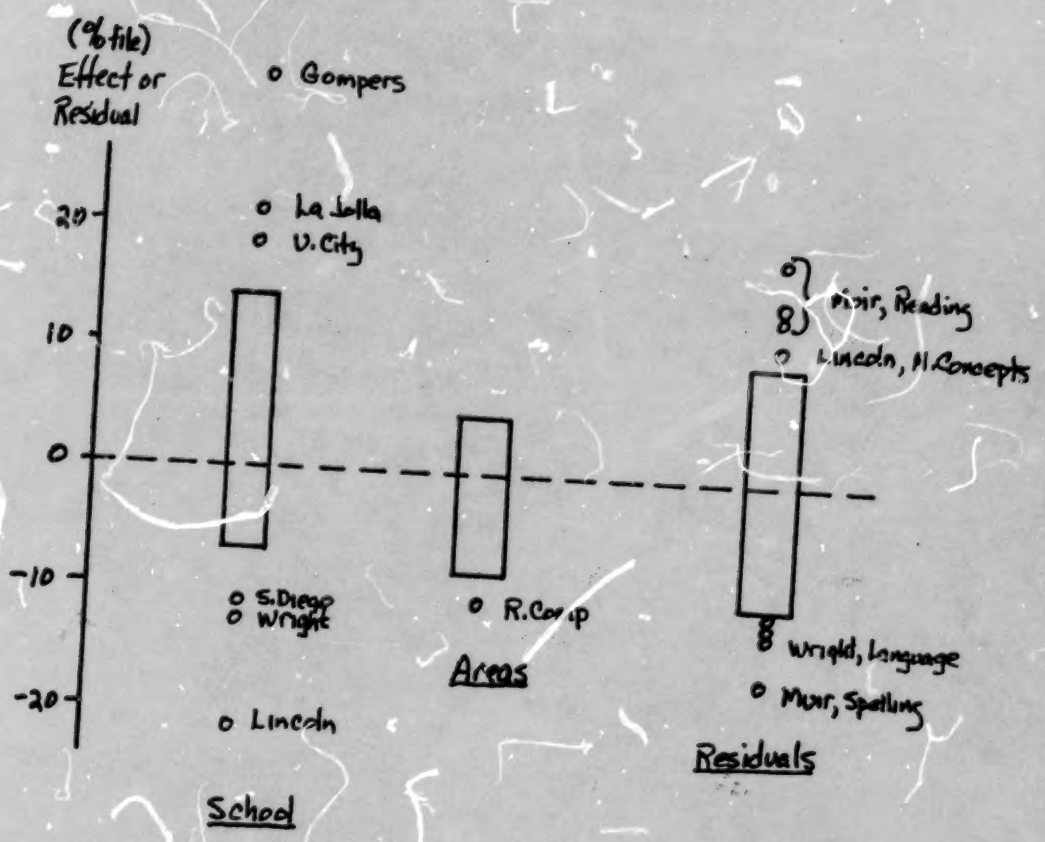
SCHOOL NAME	READING VOCAB.	READING COMPR.	TOTAL READING	SPELLING	LANG. MECH.	LANG. EXPR.	TOTAL LANGUAGE	MATH COMPUT.	MATH CONC./APPL.	TOTAL MATH
DISTRICT	50	46	50	63	58	57	59	62	[illegible]	64
CLAIREMONT	44	34	41	50	54	40	[illegible]5	53	43	45
CRAWFORD	47	46	46	63	58	60	61	64	60	64
CREATIVE AND PERFORMING ARTS	54	41	51	51	48	53	51	55	39	48
GOMPERS	86	77	82	78	74	80	82	91	87	88
HENRY	54	[illegible]	58	69	62	66	69	64	65	70
HOOVER	34	30	29	50	49	35	42	60	53	55
[illegible]	41	34	37	63	54	45	51	62	60	64
LA JOLLA	68	67	70	74	67	75	73	74	75	78
LINCOLN	18	10	14	42	38	13	25	44	43	34
MADISON	52	50	52	69	58	53	[illegible]6	64	60	64
MIRA MESA JR/SR	55	57	58	69	62	64	65	62	65	66
MISSION BAY	60	50	57	63	54	53	56	64	60	64
MORSE	38	30	33	58	49	40	48	57	[illegible]3	55
MUIR ALTERNATIVE	60	50	54	34	52	40	44	46	[illegible]3	40
POINT LOMA	57	50	57	63	58	57	61	64	62	68
SAN DIEGO	34	25	24	50	49	31	40	53	35	39
SERRA JR/SR	50	50	50	63	58	57	60	64	60	64
UNIVERSITY CITY	65	63	67	69	62	70	68	71	74	76
WRIGHT BROTHERS	30	25	31	50	38	26	23	47	43	45

BEST COPY AVAILABLE

Districtwide Testing Results

School Effects, Area Effects, and Residuals

School	Math			Language				Reading			School Effect
	Comp	Concept	Total	Spell	L. Mech.	L. Expr.	Total	Voc	Comp	Total	
Gompers	2	1	0	-10	-10	0	-1	7	4	6	33
La Jolla	-4	0	-4	-3	-6	6	1	0	5	5	22
University City	-4	2	2	-5	-8	4	-1	0	4	5	19
Henry	-6	2	1	0	-3	5	5	-6	7	1	14
Mira Mesa	-6	0	1	2	-1	5	3	-3	5	3	12
Point Loma	-2	-1	3	-2	-3	0	1	1	0	4	10
Madison	0	-1	1	6	-1	-2	-2	-2	2	1	8
Mission Bay	0	-1	1	0	-5	-2	-2	6	2	6	8
Serra	0	-1	1	0	-1	2	2	-4	2	-1	8
Crawford	1	0	2	1	0	6	4	-6	-1	-4	7
Kearny	4	5	7	6	2	-4	-1	-7	-8	-8	2
C & P Arts	1	-12	-5	-2	-1	8	3	10	3	10	-2
Morse	5	4	4	7	2	-3	2	-4	-6	-6	-4
Muir	-6	-6	-11	-17	5	-3	-2	13	14	15	-4
Clairemont	2	-5	-5	0	8	-2	0	3	-1	3	-5
Hoover	-2	6	6	1	4	-6	-2	-6	-4	-8	-6
San Diego	9	-6	-4	7	10	-4	2	0	-3	-7	-12
Wright	5	4	4	9	1	-7	-13	-2	-1	2	-14
Lincoln	10	12	1	9	9	-12	-3	-6	-8	-7	-22
Effects	6	3	5	5	1	-3	0	-4	-10	-7	50



SCHOOL EFFECTS, TEST AREA EFFECTS, AND RESIDUALS FROM 2-WAY ANALYSIS

scores about twenty-two percentile points below the national average. We have reordered the schools according to school effects. We have also ordered the subject areas by effects. The district scores on the math tests are higher, in general, than those on the language tests, which are higher than those on reading. Scores on total math tend to be about twelve percentile points higher than scores on total reading. Language scores are mixed. The district does relatively well in spelling and not as well in language expression.

The residuals also provide information about schools. For example, Gompers is a high achieving school and Lincoln is a low achieving school. Yet, scores for Gompers in spelling and language mechanics are about ten percentile points less than we would predict on the basis of knowing the school and subject effects. Scores in total reading are about six percentile points more than we would predict on the basis of school and subject effects. Relative to our model we could say that Gompers is relatively strong in reading and weak in spelling and language mechanics.

Lincoln is a low achieving school. They score low on all tests. Yet, their math scores are above our predictions based on school and subject effects. Their reading scores are below

our predictions. To clarify things a little better perhaps, we would say that the total math score for Lincoln (34) is one percentile point better than our prediction (50 - 22 + 5 = 33). The residuals become a type of "diagnostic" report on the test. Positive residuals are indicators of performance above the prediction. Negative residuals are indicators of performance below the prediction.

Discussion

This paper is really a how-to-do paper with some examples that illustrate the utility of the process. For the most part, the methods come direct from Tukey's truely monumental book Exploratory Data Analysis. I have added only the attempt to extract cohort effects from district test data, primarily because school people believe in these effects.

Most school district test reports contain endless pages of test data. The amount of information is staggering for even relatively small districts. Most of the test data is organized into some form of two-way table. Sheets of paper are essentially two-dimensional and reports take advantage of that by organizing data along two dimensions. Our aim is to simplify, however. How can we summarize the vast information provided? We seek to gain insight into our data by summarizing, tearing apart two-dimensional data into component pieces that are grist for our conversations about schools and schooling.

Underlying this paper is more than just another methodology for manipulating test data. Underlying this paper is a philosophy and attitude about how we should analyze data.

Today is the day of the computer. We read that unless we use the computer in our daily lives to the maximum extent possible, we will not be productive and be lost in a job market that emphasizes computer skills. With the advent of the personal computer, computing has become available to all. The growth in mainframe techniques, approaches and packages has also been remarkable. Technology in computing has been in a revolutionary phase.

Yet, the analyses presented in this paper were performed with only pencil and paper. All calculations were performed either in the analyst's mind or on scratch paper. Tukey's whole approach to data analysis relies primarily on the paper and pencil technology. Throw in some tables for logarithms, squares and square roots and you can do any of the analyses in his book. Some would say that this is a step backwards. It is not.

Much in the way of data analysis can be done before we reach the computer. For example, in consulting with districts you often have only a few hours in which to conduct your analyses, not nearly enough time to obtain existing data files or key in

data to new files. Analyses need to be performed on the spot. There is just not the time to devote to preparing, editing and analyzing data sets while on-site. Besides, school districts are more interested in discussing the consequences of the analyses than they are in discussing the process of the analyses. In addition, there is that one characteristic of using analysis packages with which all analysts must live. As powerful as systems are, often we cannot get what we want.

There is one other benefit to hand calculation in data analysis situations. Mundane tasks such as transcribing, grouping and calculating enable us to become familiar with our data. These tasks give us clues as to what paths will provide fruitful investigations. It gives our minds time to sort things out and synthesize.

This is not to say that the computer is bad. For large data sets the computer is indispensable. Obtaining test scores and building statistics requires a computer for most school districts. Performing two-way analyses for many data sets where means rather than medians are removed is best performed by the computer. Indeed, much of the understanding of how well the simple methods used in data analysis work is based on experimentation by computer.